Remarks/Arguments

Specification and drawings have not been amended. Claim 20 has been amended. No new claims have been added. No claims have been cancelled. Claims 1-29 remain pending in this application. Reexamination and reconsideration of the application as amended are respectfully requested.

Claim Rejections - 35 U.S.C. § 101

The Examiner rejected Claim 20 stating: "claiming software code portions for performing method according to Claim 1 is non-statutory". To overcome this rejection, Claim 20 is now amended to replace the phrase "software code portions" with the phrase "computer-executable instructions". Please see the Listing of Claims on page 2.

Claim Rejections - 35 U.S.C. § 102

The Examiner rejected Claims 1-6, 10-15, and 19-26 under 35 U.S.C. § 102(b) as being anticipated by Schmitt et al. (US 5,062,143; hereinafter "Schmitt"). Applicant respectfully traverses this rejection for the reasons set forth below.

Anticipation requires that the "four corners" of the reference include each and every limitation of a given claim. The examiner contends, incorrectly, that Schmitt expressly discloses every limitation in each of Claims 1-6, 10-15, and 19-26. They do not.

Schmitt teaches a mechanism for examining a body of text and identifying its language by comparing successive strings of characters (i.e. trigrams) into which that body of text is parsed, with a library of key sets of trigrams, one for each language (see Schmitt col. 1, lines 53-56). Each trigram comprises the contents of three successive character/space positions in the body of text; at least some of the trigrams overlapping adjacent words (see Schmitt col. 9, lines 41-44). For example, the string of text "THIS SENTENCE IS WRITTEN IN ENGLISH" with a total of N=34 character/space positions, is parsed into all of its possible sequentially occurring trigrams producing a sequence of (N – 2)=32 trigrams comprising THI, HIS, IS_, S_S, SE, ... GLI, LIS, ISH. Adjacent trigrams such as THI and HIS overlap by two characters, while some of the trigrams such as S_S, E_I, S_W, N_I and N_E overlap adjacent words. Each of these trigrams is compared with a set of trigrams that make up the identification database of a particular language (e.g. English, French, German, Turkish, etc.). For each language there is unique key set of trigrams that have been predetermined to posses a prescribed frequency of occurrence. For example, the trigram _TH is one of the trigrams that is prevalent in English, while the trigram IE_ is one of those that occurs frequently in German text. Thus, whether or not the body of text under investigation is written in a particular language can be determined by observing the percentage of time that a parsed trigram finds a match in the language-specific key set of trigrams database (see col. 3, lines 38-65). This is different from Applicant's claimed invention, which provides a method for automatically filtering a corpus of documents containing textual and non-textual information. Textual information is commonly formatted for the human eye, intermingled with non-textual information such as tables, graphics, images, etc. When such textual information needs to be processed by a machine (e.g. for delivery to a human through speech synthesis or for language translation and for statistical analysis purposes), it becomes necessary to separate what really constitutes text (i.e. a succession of words

and punctuation) from the non-textual information. The term "filtering" is referred in the specification to the removing from the collection of documents making up the corpus, those portions which are not representative of (i.e. are not in conformity with) the language under consideration, such as non-textual portions and textual portions expressed in another language (see Application page 7, lines 3-11). According to the Applicant's invention, the corpus of documents is divided into appropriate portions (e.g. lines, paragraphs or whole documents) and each portion is treated independently from the other portions (see Application page 7, lines 17-26). For each portion of the corpus of documents, a regularity value is determined and then compared with a threshold value to determine whether the portion's conformity is sufficient or not. The portions whose conformity is not sufficient are rejected from the corpus of documents (see Application page 10, lines 26-30 and page 11, lines 12-14). These distinctions, which are clearly recited in each and every one of the independent claims, are apparent in that Applicant's claimed invention provides for filtering of undesired portions of the corpus of documents, which is different from Schmitt's determination of the language in which a body of text is written. Applicant's invention is solving a different problem than the Schmitt's invention. For this reason alone, Schmitt cannot anticipate the subject matter of Applicant's claimed invention and the § 102(b) rejection of the Claims 1-6, 10-15, and 19-26 cannot stand.

Claims 1, 10 and 21

Schmitt col. 3, lines 24-26 recite:

> *"...compare the trigrams into which the input body of text is parsed with contents (approximately 80 trigrams) of each trigram key set in the library."*

Schmitt figure 3, block 301 recites:

"PARSE TEXT INTO SUCCESSIVE TRIGRAMS"

The Examiner states that Schmitt col. 3, lines 24-25 and figure 3, block 301, as recited above, "shows parsing text which is same as dividing the corpus of documents into portions". Applicant disagrees and believes that the Examiner has misconstrued the reference, as Schmitt's parsing of the text into successive trigrams for the purpose of determining the language of the text is not the same or equivalent to Applicant's dividing the corpus of documents into appropriate portions for the purpose of filtering (i.e. removing) non-conforming portions. Although the acts of parsing and dividing of a document into pieces may be similar, the mechanisms by which the document is parsed or divided into pieces and the sizes and the nature of the pieces that the document is parsed or divided into are quite different. The dividing mechanism and portions in Applicant's invention are not the same or equivalent to parsing mechanism and trigrams in Schmitt's invention.

According to Schmitt, the body of input text is specifically parsed into all of its possible sequentially occurring strings of characters called trigrams. Each trigram comprises the contents of three successive character/space positions of the body of text, overlapping with next trigram in the sequence by two characters. At least some of the trigrams overlap adjacent words. It is important that the body of text is parsed into trigrams only, because they will be compared with key sets of trigrams that have been previously determined to identify and be uniquely associated with respectively different languages such as English, French, German, Turkish, etc. (see Schmitt col. 3, lines 38-65, col. 6, lines 1-5, and col. 9, lines 41-44).

According to Applicant's invention, the corpus of documents are not parsed or divided into trigrams (i.e. three successive characters/spaces, some of which overlap adjacent words). Instead, they are divided into appropriate portions such as lines, paragraphs or whole documents whose size is determined as a function of the document corpus's overall size and/or as a function of the nature of the documents contained in the corpus. For example, for a corpus of documents written in English, the portion size suitable for the dividing step would be a paragraph. With a paragraph being defined as a set of characters that is isolated upwards and downwards by at least one blank line. Each portion resulting from the dividing step is treated independently (see Application page 7, lines 17-26).

Schmitt col. 3, lines 26-32 recite:

> *"If the percentage of parsed trigrams in the body of text for which copies are found in a respective key set is at least equal to a preselected value, based upon a previously conducted probability of occurrence determination, then the language of that key set is chosen as one possible language in which the body of text is written."*

The Examiner states that Schmitt col. 3, lines 26-32, as recited above, teaches determining for each portion of the corpus of the documents a regularity value measuring the conformity of the portion with respect to character sequences probabilities predetermined for the language. Again Applicant disagrees and believes that the Examiner has misconstrued the reference, as Schmitt's percentage of parsed trigrams found in a language database (i.e. key set of trigrams) is not the same or equivalent to Applicant's regularity value for each portion. The regularity value of a portion in Applicant's invention has a specific meaning and is calculated quite differently than the Schmitt's

simple calculation of the percentage of parsed trigrams found in a database. The calculated percentage in Schmitt's invention applies to a collection of trigrams found in the database, not to an individual trigram in the collection. The calculated regularity value in Applicant's invention on the other hand is unique to each individual portion and does not apply to a collection of portions.

According to Schmitt, when all (N – 2) trigrams have been compared with the contents of the language database (i.e. key set of trigrams), the percentage of matches or hits found in the database is determined from the ratio of the total hits to the number (N –2) of trigrams. A software counter is incremented each time a trigram is found in the database to maintain the total number of hits. The percentage of matches or hits found applies to a collection of trigrams not to any particular trigram (see Schmitt col. 4, lines 1-13 and figure 3, blocks 303, 304, 305, 306 and 311).

According to Applicant's invention, the regularity value V_R for each portion is based on a computed perplexity of the portion with respect to a language statistical model, for example, a well known in the art character-based N-gram model. In general, a language statistical model such as the N-gram model, tries to predict the current character based on the preceding N characters. It is suggested to compute the perplexity of the orthographic representation of a word with respect to a character-based N-gram model. Perplexity is an information theory measurement, expressed as a number not percentage. It is an indication of how many different letters are likely to follow a particular context of string characters. Informally, perplexity may be regarded as the average number of following characters that a character-based language model may have to choose from, given the present history of characters already looked at. Formally, the perplexity is the reciprocal of the geometric average of the probabilities of a hypothesized string of characters. The regularity

value is calculated for each portion individually and separately and is intended to measure the conformity of each portion with respect to character sequences probabilities predetermined for the language under consideration. The calculated regularity value only applies to a given portion and does not apply to a collection of portions (see Application page 9, line 12 thru page 10, line 24).

The Examiner further states that Schmitt col. 3, lines 26-32, as recited above, teaches comparing each regularity value with a threshold value V_T to decide whether the conformity is sufficient. Again Applicant disagrees and believes that the Examiner has misconstrued the reference, as Schmitt's comparison of the percentage of hits to a prescribed threshold value is not the same or equivalent to Applicant's comparison of each and every one of the regularity values with a threshold value. Although both Schmitt and the Applicant's invention perform comparing of two things, the things they are comparing are quite different from each other. The regularity value V_R and the setting of a threshold value V_T in Applicant's invention are different from the percentage of hits and setting of the threshold value in Schmitt respectively.

According to Schmitt, the percentage of hits (i.e. ratio of the total hits to the number N – 2 of trigrams) is compared to a prescribed threshold value, associated with a selected minimum percentage of hits, (e.g. 10%). In Schmitt, when setting the threshold value, some prescribed noise margin may be subtracted from the likelihood percentage used to assemble the trigram database for a respective language. Thus, for the current example of English text, the threshold might be set at 0.20, while for a German database, the value might be somewhat lower, for example on the order of 0.10-0.18 (see Schmitt col. 4, lines 14-16 and col. 5, lines5-13).

According to Applicant's invention, and as stated earlier, the regularity value V_R for each portion is based on a computed perplexity of the portion with respect to a language statistical model. The threshold value V_T on the other hand, is determined beforehand by first defining a test corpus as a subset of the document corpus to be filtered. Then a manual cleaning is performed on the test corpus so as to obtain a cleaned test corpus, which is representative of the type of textual information that is considered as being sufficiently in conformity with the language rules. After that, a perplexity value of the cleaned test corpus with regard to the statistical model is computed. Similarly, a perplexity value of the rejected test corpus (i.e., the set of portions rejected from the initial test corpus) is also computed. Finally, the threshold value is determined between the two perplexity values obtained, for example as the average value of these two perplexity values (see Application page 11, lines 1-10).

The Examiner further states that Schmitt col. 3, lines 26-29, as recited above, teaches rejecting any portion of the corpus of documents whose conformity is not sufficient. Again Applicant disagrees and believes that the Examiner has misconstrued the reference, as Schmitt does not reject or remove any of the parsed trigrams or the body of text based on the percentage of hits falling below the preselected threshold value. Unlike Schmitt, in Applicant's invention, those portions of the corpus of documents whose conformity is determined not to be sufficient are removed (i.e. filtered) from the corpus.

According to Schmitt, if the ratio of the number of hits in the examined body of text to its total number of parsed trigrams is at least equal to the threshold value, then the text is identified as being possibly written in the language associated with that respective key set. If the ratio is less than

the threshold, then it is determined that the text is not written in that language (see Schmitt col. 4, lines 23-29). The Examiner has misconstrued this and previous references to conclude incorrectly that "... if the percentage [regularity value] of the parsed trigrams in the body of the text for which copies are in a respective key set is below preselected value then the body of text will be rejected". As matter of fact, the body of text in Schmitt never gets rejected; only a language under process may be rejected as possible candidate for the body of text to be written in. The process of language identification and elimination continues with the next language using the same parsed trigrams of the body of text. According to Schmitt, the process is repeated using each of the language databases until all the languages have been processed. After all languages have been processed, the language in the set of possible languages with the highest 'hit' or 'match' percentage, is selected and identified as the language in which the text is written. If no language was identified as a possible language, then no language identification is made (see Schmitt col. 4, line 29-37).

According to Applicant's invention, if the conformity of the portion under consideration is determined as being sufficient, the portion is kept. Conversely, if the portion is determined as being insufficient, the portion is rejected (see Application page 11, lines 12-14).

Claims 2, 11 and 22

Schmitt col. 3, lines 26-32 recite:

> *"If the percentage of parsed trigrams in the body of text for which copies are found in a respective key set is at least equal to a preselected value, based upon a previously conducted probability of occurrence determination, then the language of that key set is chosen as one possible language in which the body of text is written."*

The Examiner states that "... a previously conducted probability of occurrence determination" in Schmitt col. 3, lines 26-32, as recited above, corresponds to "character sequences probability derived from statistical model representative of the language" in the Applicant's invention. Applicant disagrees and believes that the Examiner has misconstrued the reference, as Schmitt's probability of occurrence determination is not the same or equivalent to Applicant's character sequences probability.

According to Schmitt, the probability of occurrence determination refers to determining which trigrams (approximately 80 trigrams) best represent a particular language. "In order to determine which trigrams are to make up the database of a particular language, a reasonably sized section (e.g. 3,000-5,000 characters) of text is parsed into a plurality of trigrams. A running count is maintained of the occurrence of each of the trigrams that has been parsed from that section of text, and the ratio of each of the number of occurrences of the trigrams thus counted with the total number of trigrams into which the section of text has been parsed is calculated. From these ratios a characteristic is derived which is representative of the frequency of trigram occurrence of each trigram that may be formed using both the characters of that language and a space position. An examination of this ratio characteristic will reveal which trigrams occur at least some prescribed percentage of time (e.g. 30-35% for English) that is representative of a high likelihood of association with that language" (see Schmitt col. 4, lines 44-61).

According to Applicant's invention, the character sequences probability refers to determining which character would most likely follow a given string of characters. The character sequences

probabilities are derived from a statistical model well known in the art, such as a character-based N-gram model, representative of the language considered. In general, an N-gram model tries to predict the probability of an N character long string occurring in a given language. The occurrence of a character C is completely determined by the past N characters (see Application page 9, line 30 thru page 10, line 6).

Claims 3, 12 and 23

Schmitt col. 3, lines 26-32 recite:

> *"If the percentage of parsed trigrams in the body of text for which copies are found in a respective key set is at least equal to a preselected value, based upon a previously conducted probability of occurrence determination, then the language of that key set is chosen as one possible language in which the body of text is written."*

The Examiner states that Schmitt col. 3, lines 26-32, as recited above, discloses regularity value, which is based on a computed perplexity of the portion with respect to the statistical mode. Once again, Applicant disagrees and believes that the Examiner has misconstrued the reference. As previously explained, Schmitt's percentage of parsed trigrams found in a language database (i.e. respective key set of trigrams) is not the same or equivalent to Applicant's regularity value for each portion. The regularity value of a portion in Applicant's invention has a specific meaning and is calculated quite differently than the Schmitt's simple calculation of the percentage of parsed trigrams found in a database. The calculated percentage in Schmitt's invention applies to a collection of trigrams found in the database, not to an individual trigram in the collection. The

calculated regularity value in Applicant's invention on the other hand is unique to each individual portion and does not apply to a collection of portions.

According to Schmitt, when all (N – 2) trigrams have been compared with the contents of the language database (i.e. key set of trigrams), the percentage of matches or hits found in the database is determined from the ratio of the total hits to the number (N –2) of trigrams. A software counter is incremented each time a trigram is found in the database to maintain the total number of hits. The percentage of matches or hits found applies to a collection of trigrams not to any particular trigram (see Schmitt col. 4, lines 1-13 and figure 3, blocks 303, 304, 305, 306 and 311).

According to Applicant's invention, the regularity value V_R for each portion is based on a computed perplexity of the portion with respect to a language statistical model, for example, a well known in the art character-based N-gram model. In general, a language statistical model such as the N-gram model, tries to predict the current character based on the preceding N characters. It is suggested to compute the perplexity of the orthographic representation of a word with respect to a character-based N-gram model. Perplexity is an information theory measurement, expressed as a number not percentage. It is an indication of how many different letters are likely to follow a particular context of string characters. Informally, perplexity may be regarded as the average number of following characters that a character-based language model may have to choose from, given the present history of characters already looked at. Formally, the perplexity is the reciprocal of the geometric average of the probabilities of a hypothesized string of characters. The regularity value is calculated for each portion individually and separately and is intended to measure the conformity of each portion with respect to character sequences probabilities predetermined for the

language under consideration. The calculated regularity value only applies to a given portion and does not apply to a collection of portions (see Application page 9, line 12 thru page 10, line 24).

Claims 4, 13 and 24

Schmitt col. 3, lines 16-20 recite:

> *"As described briefly above, the processing methodology employed by the present invention uses a key set, the collection of key sets for all languages in the language constituting a library, of trigrams, each set being associated with a respectively different language."*

The Examiner states that Schmitt col. 3, lines 16-18, as recited above, discloses a statistical model previously elaborated from a reference document determined as conforming with the rules of the language, which corresponds to Schmitt's library of key sets of all languages. Once again, Applicant disagrees and believes that the Examiner has misconstrued the reference, as Schmitt's library of key sets of all languages does not correspond, and is not the same or equivalent to Applicant's statistical model. Although both Schmitt and Applicant use a reference document to elaborate the library of key sets of trigrams and the statistical model respectively, they use the reference document in a different way and for different purposes with different results.

According to Schmitt, and as recited above, the collection of key sets for all languages constitutes a library of trigrams, each set of trigrams being associated with a different language. Each set of trigrams for a given language comprises approximately 80 trigrams. To determine which trigrams are to make up the key set for a particular language, a reasonably sized section of text (i.e. a reference document) is parsed into a plurality of trigrams and the probability of occurrence of each

trigram is determined. Those trigrams that occur at least some prescribed percentage of time (e.g. 30-35% for English) are selected as representative of a high likelihood of association with that language (see also Schmitt col. 4, line 57 thru col. 5, line 3).

According to Applicant's invention, the statistical model is a character-based N-gram model. Character-based N-gram models are well known in the art and do not constitute a library of key sets of trigrams. In general terms, a language model, as for instance an N-gram model, tries to predict the a-priori probability of an N character long string occurring in a given language (see Application page 9, line 27 thru page 10, line 1). To elaborate a statistical model for a given language, a corpus of textual data (i.e. a reference document) conforming to the rules of that language (i.e. types of word, punctuations, line breaks, special characters, etc.) is first collected. The collection of textual data obtained is then manually cleaned to keep only pertinent textual data (e.g. graphics, tables, images and other language text are suppressed). A clean training corpus is therefore obtained and stored. The clean training corpus is then subdivided into training data and held-out data, by randomly selecting a certain percentage of the corpus (e.g. 10%). The training data is served as a basis to compute N-grams statistics upon which the language statistical model is determined. The held-out data is used to optimize the statistical model computed from the training data. The training data is used to compute 1-gram, 2-gram and 3-gram models. The models are computed by counting uni-letter frequencies, bi-letter frequencies, and tri-letter frequencies. The frequencies obtained are then used as approximations of the probability of such letter sequences. The construction and functioning of such N-grams models is known within the state of the art. The overall likelihood of a sequence of 3 letters is computed as a linear combination of the uni-letter, bi-letter and tri-letter likelihood, with an added offset to give non-zero probabilities to never-observed letter sequences.

The coefficients of the linear combination are estimated using the held-out data in order to optimize the performance of the statistical model. The final statistical model is generated and stored (see Application page 11, line 29 thru page 12, line 28).

Claims 5, 14 and 25

Schmitt col. 3, lines 16-20 recite:

> *"As described briefly above, the processing methodology employed by the present invention uses a key set, the collection of key sets for all languages in the language constituting a library, of trigrams, each set being associated with a respectively different language."*

Schmitt col. 3, lines 30-31 recite:

> *"... probability of occurrence determination, then the language of that key set is chosen as one possible language ..."*

The Examiner states that Schmitt col. 3, lines 16-19 and col. 3 lines 30-31, as recited above, disclose statistical model being determined according to N-gram statistics. Once again, Applicant disagrees and believes that the Examiner has misconstrued the references. As previously explained, Schmitt's library of trigrams and probability of occurrence do not correspond to Applicant's statistical model and N-gram statistics (i.e. character sequences probabilities) respectively. Please see the explanation above under Claims 2, 11, 22 and under Claims 4, 13, 24.

Claims 6, 15 and 26

Schmitt col. 3, lines 16-20 recite:

> *"As described briefly above, the processing methodology employed by the present invention uses a key set, the collection of key sets for all languages in the language constituting a library, of trigrams, each set being associated with a respectively different language."*

Schmitt col. 3, lines 30-31 recite:

> *"... probability of occurrence determination, then the language of that key set is chosen as one possible language ..."*

The Examiner states that Schmitt col. 3, lines 16-19 and col. 3 lines 30-31, as recited above, disclose statistical model is character-based N-gram model. Once again, Applicant disagrees and believes that the Examiner has misconstrued the references. As previously explained, Schmitt's library of trigrams and probability of occurrence do not correspond to Applicant's statistical model and N-gram model respectively. Please see the explanation above under Claims 2, 11, 22, under Claims 4, 13, 24 and under Claims 5, 14, 25.

Claims 19, 20

Schmitt figure 1, blocks 20 and 30:

> *"ENCODING DIGITIZER" and "MEMORY/PROCESSING UNIT"*

The Examiner states that Schmitt's figure 1, blocks 20 and 30, as recited above discloses computer system and software code. Although both Schmitt and Applicant disclose a computer system and software code, Applicant respectfully points out that the computer system and software code in Schmitt is for language identification (e.g. English or French) of an input text, which is quite different than Applicant's computer system and software code for filtering (i.e. elimination of non-textual data such as graphics, tables, images, ...) a corpus of documents using statistical models well known in the art.

Allowable Subject Matter

The Examiner objected to Claims 7-9, 16-18 and 27-29 as being dependent on rejected base claim but otherwise allowable if written in independent form including limitation of the base claim and any intervening claims. Applicant respectfully believes that based on the detailed explanation given above for each and every one of the claims, there is no need to re-write these claims in independent form. Applicant respectfully believes that the claims including the amended Claim 20 as listed in the Listing of Claims section of this paper are now in condition for allowance.

Conclusion

Applicant therefore respectfully requests that the Examiner reconsider all currently outstanding objections and rejections and that they be withdrawn. It is believed that a full and complete response has been made to the outstanding Office Action and, as such, the present application is in condition for allowance. If the Examiner believes, for any reason, that personal communication will expedite prosecution of this Application, the Examiner is invited to telephone the undersigned at the number provided. Prompt and favorable consideration of this Response is hereby solicited.

Respectfully submitted,

Hubert Crepy

By: _____

Farrokh E. Pourmirzaie, Reg. No. 45,297
Agent for Applicant
International Business Machines Corporation
Intellectual Property Law
555 Bailey Avenue, J46A/G462
San Jose, CA 95141-9989
Telephone: (408) 463-3539

Date: August 24, 2004